CHAPTER 2 _ MACHINE LEARNING LIFE CYCLE

Machine Learning learns from data. It understand the pattern of data and makes future prediction based on pattern. Their are several steps involved in order to solve business problem using Machine Learning.

Machine Learning life cycle is a process that helps to build effective Machine Learning models.This process involves preparation of data related to problem, preprocessing data, training model and deploying into production.

Machine Learning Life cycle has 6 major steps :

1. Problem Definition
2. Data Selection or Data Gathering
3. Exploratory Data Analysis
4. Data Preprocessing
5. Model Selection, Model Training and Model Evaluation
6. Model Deployment

1. Problem Definition

Initially we need to understand business problem. Let us take one example XYZ bank their is incresing the losses every year by giving loans. The problem here is to find whether loans should be given or not based on user details. This is classification problem.

2. Data Selection or Data Gathering

In this step is to find data related to this problem. Data is collected to various data sources such as files, database, internet or mobile device. This is most important step in Machine Learning life cycle. The quality and quantity of collected data will determine efficiency of model.

3. Exploratory Data Analysis

In this step we will understand about data and we will get important insights of data. Let's take an example Loan prediction dataset.

The training dataset contains 614 rows and 13 columns out of which 1 is dependent are other 12 are independent variables .

A sample from the dataset is mentioned below:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	A
0	LP001002	Male	No	0	Graduate	No	
1	LP001003	Male	Yes	1	Graduate	No	
2	LP001005	Male	Yes	0	Graduate	Yes	
3	LP001006	Male	Yes	0	Not Graduate	No	
4	LP001008	Male	No	0	Graduate	No	

Property_Area	Loan_Status
Urban	Y
Rural	N
Urban	Y
Urban	Y
Urban	Y

A small description of each column is given below:

- Loan_ID:Unique Loan ID
- Gender:Male/ Female
- Married:Applicant married (Y/N)
- Dependents:Number of dependents
- Education:Applicant Education (Graduate/ Under Graduate)
- Self_Employed:Self employed (Y/N)
- ApplicantIncome:Applicant income
- CoapplicantIncome:Coapplicant income

- LoanAmount:Loan amount in thousands
- Loan_Amount_Term:Term of loan in months
- Credit_History:credit history meets guidelines
- Property_Area:Urban/ Semi Urban/ Rural
- Loan_Status:Loan approved (Y/N)

We will do summarization on data

	Loan_ID	Gender	Married	Dependents	Education	Self_Employ
count	614	601	611	599	614	5
unique	614	2	2	4	2	
top	LP001421	Male	Yes	0	Graduate	
freq	1	489	398	345	480	5
mean	NaN	NaN	NaN	NaN	NaN	N
std	NaN	NaN	NaN	NaN	NaN	N
min	NaN	NaN	NaN	NaN	NaN	N
25%	NaN	NaN	NaN	NaN	NaN	N
50%	NaN	NaN	NaN	NaN	NaN	N
75%	NaN	NaN	NaN	NaN	NaN	N
max	NaN	NaN	NaN	NaN	NaN	N

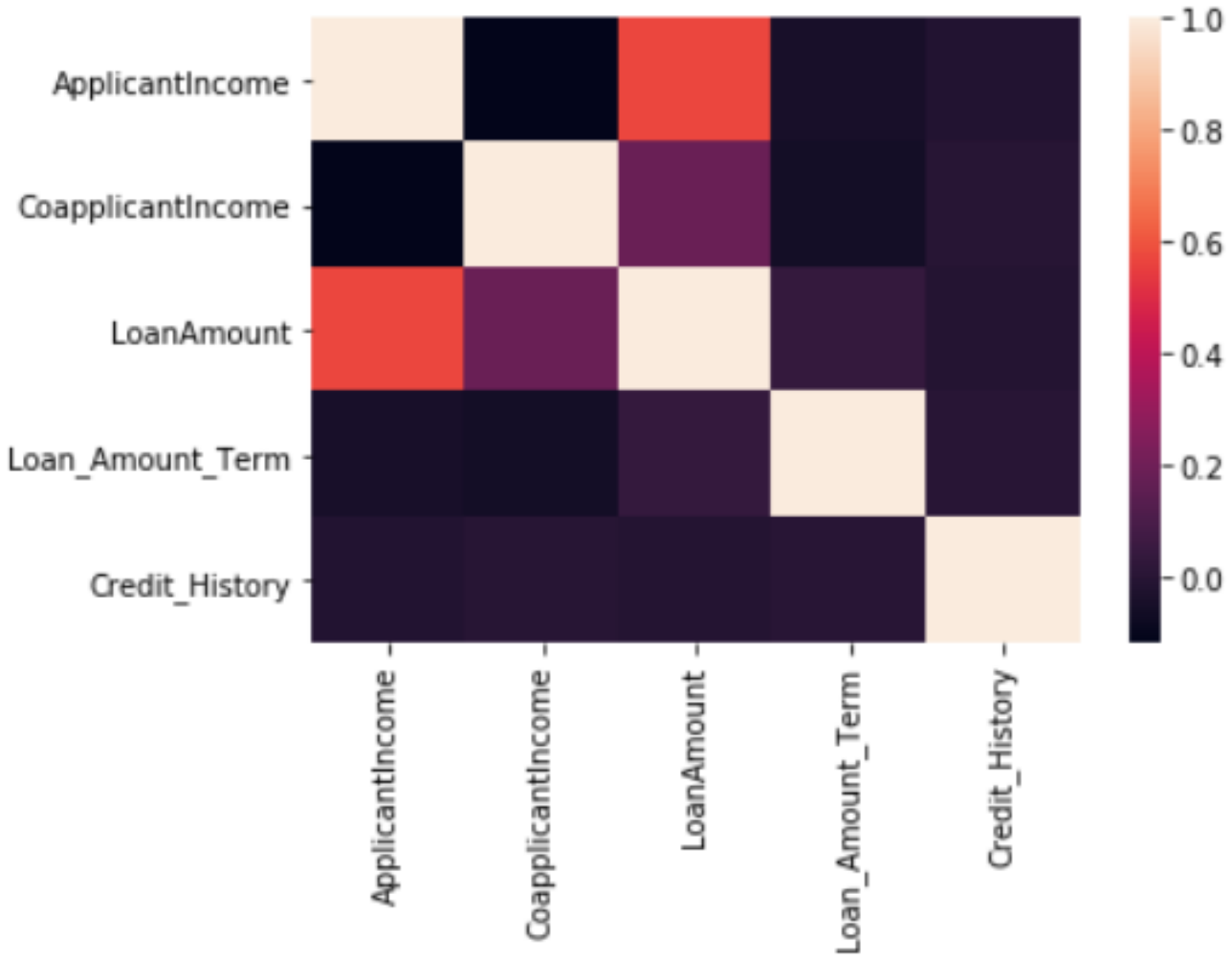
Credit_History	Property_Area	Loan_Status
564.000000	614	614
NaN	3	2
NaN	Semiurban	Y
NaN	233	422
0.842199	NaN	NaN
0.364878	NaN	NaN
0.000000	NaN	NaN
1.000000	NaN	NaN
1.000000	NaN	NaN
1.000000	NaN	NaN
1.000000	NaN	NaN

- Data has integer, float and object values
- Data has null values in Dependents, Self_Employed , LoanAmount, Loan_Amount_Term and Credit_History column as count is not equal to the number of rows
- The numerical operations are returned as NaN for the object type columns

```
train['Loan_Status'].value_counts()
```

```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

- It shows the unique value counts in the target variable i.e. Y or N
- It shows that out of 614 data in given dataset, 422 Y and 192 N that means this dataset is slightly imbalanced we need to make balanced.



- This is the correlation matrix for the numerical columns of the data.
- ApplicantIncome and LoanAmount are highly co-related.

4. Data Preprocessing

Data Preprocessing or Data Wrangling is the process of cleaning and converting data into usable format.

- **Data cleaning**

  In this process we will treat Null values, Outliers, Duplicate data

- **Data Transformation**

  In this process we will try to convert categorical data into numerical data. And we will find out whether we can extract some useful insights from the existing columns and we will transform the columns into certain ways which will be useful for model.

- **Feature Selection**

  Feature Selection is the process where you automatically or manually select those

features which contribute most to your prediction variable or output in which you are interested in.

Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

5. Model Selection, Model Training and Model Evaluation

Here, for model selection, we have taken various ensemble classifier such as decision tree, random forest, bagging, xgboost etc. and we trained each one of the model and made a comparison between the evaluation metrics of each of them. For evaluation metrics, we have taken accuracy, precision and recall.

```
eval_df
```

	Classifier	Precision	Recall	Accuracy
0	Booster	0.827366	0.792986	0.815642
1	XGBClassifier	0.845624	0.793758	0.821229
2	XGBRFClassifier	0.840236	0.804505	0.826816
3	DecisionTreeClassifier_gini	0.800667	0.761197	0.787709
4	DecisionTreeClassifier_entropy	0.800667	0.761197	0.787709
5	RandomForestClassifier	0.800667	0.761197	0.787709
6	Bagging_BaggingClassifier	0.800667	0.761197	0.787709
7	Bagging_BaggingClassifier	0.800667	0.761197	0.787709
8	Bagging_BaggingClassifier	0.800667	0.761197	0.787709
9	Bagging_BaggingClassifier	0.800667	0.761197	0.787709
10	Bagging_BaggingClassifier	0.800667	0.761197	0.787709
11	AdaBoostClassifier	0.800667	0.761197	0.787709
12	GradientBoostingClassifier	0.800667	0.761197	0.787709

From the above comparison, we found out that XGBoostRandomForest classifier works

best on the loan prediction dataset here with accuracy of 82.7%, precision of 84.0% and recall of 80.0%.

6. Model Deployment

Last step of Machine Learning Life cycle is to deploy effective model for production. Model will be acceptable only if it gives accurate results, with acceptable speed.